The CDF Run 2 Offline Computer Farms

Jaroslav Antos[1,2], Marian Babik[1], Yen-Chu Chen[2], Troy Dawson[3], Lisa Giacchetti[3], Terry Jones[3], Tanya Levshina[3], Igor Mandrichenko[3], Ramon Pasetes[3], Karen Shepelak[3], Miroslav Siket[2,4], Steven Timm[3], Stephen Wolbers[3], G.P. Yeh[3], Ping Yeh[5]

(1) *Institute of Experimental Physics, Slovak Academy of Sciences, Slovak Republic*
(2) *Institute of Physics, Academia Sinica, Taiwan, ROC*
(3) *Fermi National Accelerator Laboratory, Batavia, IL, USA*
(4) *Comenius University, Slovak Republic*
(5) *Department of Physics, National Taiwan University, Taipei, Taiwan, ROC*

Abstract

Run 2 at Fermilab began in March, 2001. CDF will collect data at a maximum rate of 20 MByte/sec during the run. The offline reconstruction of this data must keep up with the data taking rate. This reconstruction occurs on a large PC farm, which must have the capacity for quasi-real time data reconstruction, for reprocessing of some data and for generation and processing of Monte Carlo samples. In this paper we will give the design requirements for the farm, describe the hardware and software design used to meet those requirements, describe the early experiences with Run 2 data processing, and discuss future prospects for the farm, including some ideas about Run 2b processing.

Keywords: PC farms, Large-scale computing systems, Event Reconstruction

1 Introduction

CDF will collect and analyze a large amount of data during Run 2a, which will occur during the years 2001-2003. A peak output rate to mass storage of 20 MByte/sec. is expected and has already been achieved in data-taking. Factoring in detector downtime, accelerator downtime and other losses a total raw data size of hundreds of Terabytes per year is expected. The CDF offline production farms are required to process this data in quasi-real time, meaning that raw data should be reconstructed with only a short delay to allow for the determination and availability of calibrations or other necessary inputs to the production executable. Therefore, the offline production farm should be able to process at 20 MByte/sec peak. In addition the farm will be expected to reprocess some data and to generate and reconstruct Monte Carlo data.

The output of the farm is physics datasets in CDF PAD format, a summary format. This output is split into many physics datasets. The splitting operation is required to place similar physics data together on disk or tape files, allowing faster and more efficient physics analysis.

2 Architecture and Data Flow

2.1 Architecture

The CDF offline production farms have been designed to satisfy the requirements listed in the previous section. The software design is described in [1][2]. The hardware design is shown in Figure 1.

The farm consists of two I/O nodes and many (currently 154) worker nodes. One of the I/O nodes (fcdfsgi1) is an SGI O2000 with four 300 MHz processors and 1 GByte of memory. This node has access to the tapes in the CDF tape robot, and is therefore able to stage data from tape to disk and write data from disk to tape. This I/O node has over 1 TByte of staging disk for the farms and many Sony AIT-2 tapedrives. There are 3 ethernet network connections

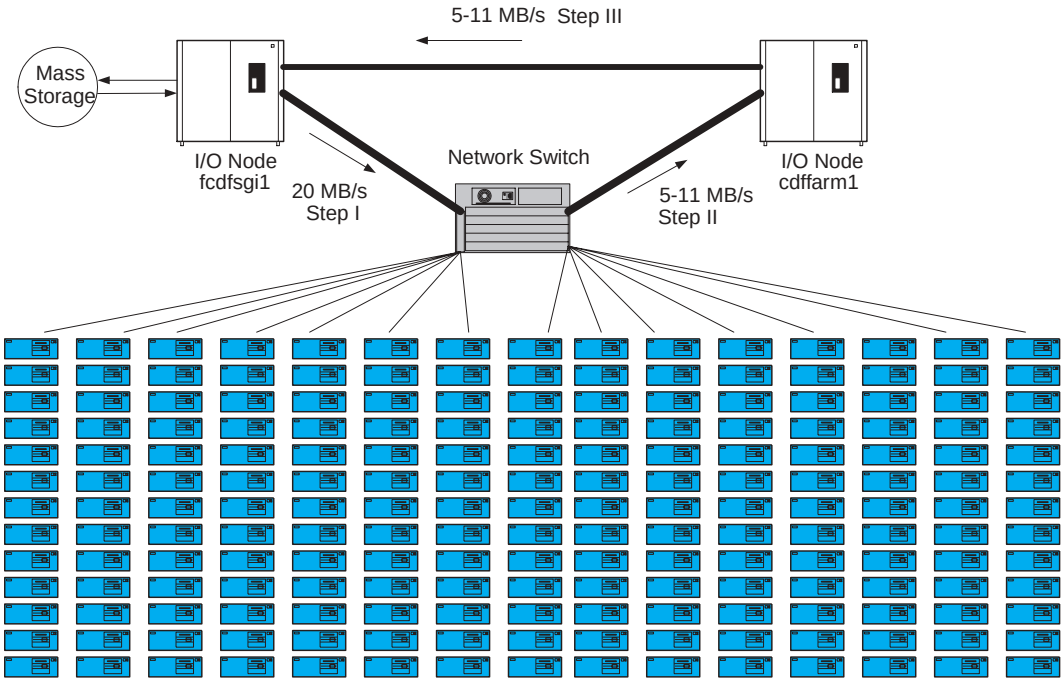

Figure 1: Design of CDF Offline Production Farms

to this node, 2 Gbit connections for the farms and one 100 Mbit connection for more general CDF traffic. The node is also used to log data to tape from the experiment.

The second I/O node (cdffarm1) is an SGI O2200 with four 400 MHz processors, 1 GByte of memory, 2 Gbit network connections (one to fcdfsgi1 and the other to the farm switch) and about 1 TByte of disk storage. This node serves as the control node for the CDF farms, as well as the NFS server for the farms.

The worker nodes consist of 50 Pentium III/500 MHz machines, 40 Pentium III/800 MHz machines and 64 Pentium III/1 GHz machines. Each PC has 2 CPU's, 512 MByte of memory and approximately 40 GByte of disk available for local storage (the operating system has its own disk). Each machine has a dedicated 100 Mbit ethernet connection to the farm switch. In this system the CPU is provided by PCs, while the I/O is handled by larger and expandable SMP systems.

The CDF farm switch is a CISCO 6509. Currently there are 2 Gbit ports and over 150 100 Mbit ports used for the CDF farm. One additional port provides an uplink to the general Fermilab backbone.

2.2 Data Flow

All raw data from the experiment (with the exception of expressline data) is first written to tape in the mass storage system. These tapes are cataloged in the CDF Data File Catalog, a set of tables in an Oracle database. After the data is written to tape and properly cataloged, and once the necessary calibration constants exist, the data is available for reconstruction on the farms. The files are staged from tape to disk on the I/O node fcdfsgi1. Each file is sent to a worker node for reconstruction and each worker node runs two simultaneous reconstruction jobs. An input file is approximately 1 GByte in size and is expected to run for about 5-6 hours on a PIII/500

MHz machine, assuming each event is about 250 KByte and takes 5 seconds to reconstruct. The output is split into multiple files, with each file corresponding to a dataset defined by the triggers. To implement this, each event is written to all datasets that are consistent with that event's triggers. Because of this requirement, an event can be written to multiple datasets and each dataset is a self-contained sample for physics analysis. CDF has approximately 8 input streams and 50 output datasets defined for Run 2.

The output files for each dataset are copied to the node cdffarm1 and are combined into larger files, with a target file size of 1 GByte. During this step, files are combined in increasing order of run section number, an integer attached to each event during data-taking. The 1 GByte files are sent to the node fcdfsgi1 for final logging to tape by the CDF Data Handling system.

The farms design envisions a peak summed data flow of about 20 MByte/sec from the first I/O node to the worker nodes, 11 MByte/sec from the worker nodes to the second I/O node and 11 MByte/sec back to the original I/O node. These data flow rates have been achieved in dedicated tests on the farms.

The expressline data is a special dataset defined to be events which satisfy a certain set of triggers and which are meant to be processed quickly. To accomplish this, these files are made available on disk on the farm I/O node fcdfsgi1 soon after being written by the online system. This is done to avoid the latency required to save enough data to fill a full tape. These files are processed through the farm as soon as they are available and then are made available on user-accessible disk.

3 Experience with Production

The CDF experiment collected data in the Tevatron commissioning run in October, 2000 and in Run 2, beginning with $p\bar{p}$ collisions ("stores") in April, 2001. In these early stores CDF collected the first significant data samples. No significant problems occurred when these events were processed through the CDF production farms. During these early runs a few days to a few weeks of data were collected. 17.5 million events were collected during these runs with a total data size of 2.5 TBytes. The average CPU/event on the farm was less than the target of 5 seconds/event. In total, 32 million events were processed (some events were processed more than once) from the commissioning run and the April, 2001 data. The total CPU time used was 45 million CPU seconds (in PIII/500 units). About 6 TBytes of data was written to mass storage.

Beginning in June, 2001, both the Tevatron and the CDF detector ran very well and began to provide significant samples for offline reconstruction. This early data was written in 4 streams and the output of the farms was split into 7 output datasets. Because of debugging and other needs, the output was not in the condensed PAD format, but consisted of raw data plus reconstructed objects. The CDF experiment wrote data at a peak rate of 20 MByte/sec, which is the design goal. The farms were able to reconstruct data at the same peak rate. However, the output systems were not able to keep up with the large output of the farms. Some adjustments to the output system were made to increase the capabilities to handle this data. More staging disk was added to provide a larger buffer for the farms output and more tapedrives were added. As the run matures, the output will be reduced and should approach the design size. By making these adjustments the farms throughput should be sufficient for Run 2.

4 Prospects and Run 2b

The Tevatron collider is scheduled to run through the middle of 2002 with only a single one-month down-time. This will provide a very large data sample for the farms to process. During this time the CDF detector will be fully commissioned and the Tevatron will provide higher

luminosity. The CDF event reconstruction program will be completed, and the complete set of data streams and data sets will be put into place and used. All of this will provide quite a challenge for the CDF farms and the CDF data handling system.

The next big challenge for offline data processing for CDF will occur when the upgrade of the Tevatron collider to Run 2b occurs in 2003/2004 . This upgrade is meant to increase the luminosity by approximately a factor of 8 compared to Run 2a. Assuming that the data taking rate scales with the luminosity the offline data rate will also increase by a factor of 8. The detector will not change substantially, so it is expected that the CPU/event and the size of each event will be approximately the same as for Run 2a.

The offline production will therefore have to handle a factor of 8 times as much data and a provide a factor of 8 times as much CPU. To do this will require some changes to the current architecture. A new architecture could include a different data flow model which might require fewer or no I/O nodes, multiple 1 Gbit links, or 10 Gbit links. The PC's will all be replaced by faster PC's, which in the time frame contemplated should easily give a factor of 8 in increased computing. The data logging and storage system are assumed to scale as well by the use of a newer, higher density, higher speed tape technology. The detailed planning for these upgraded systems will begin in 2002.

5 Conclusion

The CDF offline production farm is in place and working for the processing of Run 2a data. The initial experiences have been primarily positive and adjustments are being made and will continue to be made to handle the data coming from the CDF experiment. Once this is achieved the processing of Run 2a data should proceed uneventfully. Run 2b represents a significant challenge because of the potential of a large increase in data rate and therefore of computing power required. The normal evolution of computing hardware should make it possible to meet this challenge and to handle the Run 2b data.

References

[1] Jaroslav Antos, *et al.*, "Design and First Tests of the CDF Run 2 Farms", CHEP2000, Padova, Italy, February, 2000, accepted for publication in Computer Physics Communications.

[2] Jaroslav Antos, *et al.*, "The Farm Processing System at CDF", CHEP2001, Beijing, China, September, 2001.